Survey estimation and marginal effects in Stata

Ben Jann

University of Bern, jann@soz.unibe.ch

DAGStat 2013

March, 18–22, 2013, Freiburg

Outline

- Stata: an overview
- Survey estimation
- Predictive margins and marginal effects

Stata: an overview

- History:
  - Stata started with version 1.0 in 1985. In it's first version it way simply a package for linear regression and some data management.
  - In 1987 (Stata 1.5) commands for Anova, Logit and Probit were added.
  - In 1988, Stata 2 was released containg, now supporting graphics and some survival analysis.
  - In 1990 (Stata 2.1): possibility to add user programs (ado-files)
  - In 1991: Issue 1 of the Stata Technical Bulletin (STB) was published
  - Stata 3 was released in 1992, now with a whole range of regression models for categorical data and survival analysis. Introduction of sampling weights (pweights).
  - Stata 4 in 1995: Support for panel data.
  - Stata 5 in 1996: More on panel data and survival analysis. Introduction of survey estimation.
  - 1997: First user ado-file in the SSC achive

Stata: an overview

- History:
  - Stata 6 in 1999: web aware, generic maximum likelihood estimation, tools for time-series analysis
  - Stata 7 in 2000: more on panel data, cluster analysis
  - 2001: Stata Journal was launched
  - Stata 8 in 2003: new graphics, extended GUI, multiple missing values
  - Stata 9 in 2005: Introduction of new matrix programming language (Mata), more on survey estimation, linear mixed models, multivariate analysis
  - Stata 10 in 2007: graph editor, mixed models with binary and count responses, exact logistic regression, dynamic panel estimators, Stata/MP for parallel computing
  - Stata 11 in 2009: factor variables, margins and marginal effects, multiple imputation, GMM
  - Stata 12 in 2011: structural equation modeling, more on multiple imputation

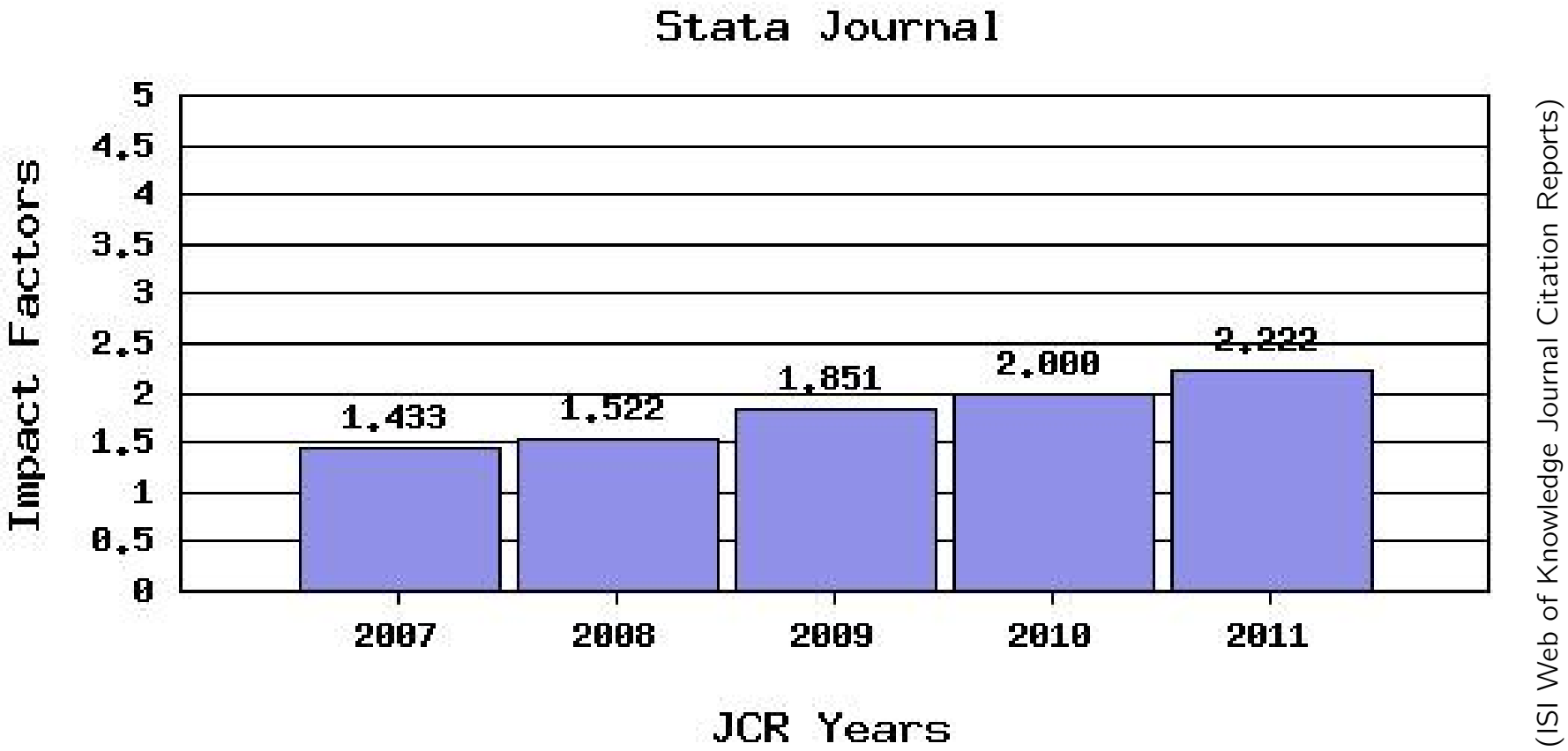
Stata: an overview

- Functionality:
  - Great tools for data management
  - All you need for basic statistics (tables, tests, nonparametric methods etc.)
  - Wide variety of linear and non-linear regression models
  - Survival analysis
  - Time series analysis
  - Panel data and multilevel/mixed-effects analysis
  - Structural equation modeling
  - Multivariate methods (factor analysis, MDS, cluster analysis etc.)
  - Resampling methods, complex survey analysis, and multiple imputation
  - Good graphics
  - Two programming languages: ado (script programming) and Mata (for serious programming)

Stata interface

Stata: an overview

- Some of the things I like about Stata:
  - Fast (most of the times), no problems with big data sets (given you have the RAM)
  - Very reliable, everything is under control, great possibilities for automation of almost anything, very scientific approach (the details matter)
  - Programming own commands is very easy, large collection of user add-ons available (SSC Archive: ideas.repec.org/s/boc/bocode.html)
  - Mata
  - Great documentation! (*example*)
  - Great books (http://www.stata.com/bookstore/books-on-stata)
  - Great technical support
  - Stata Journal (http://www.stata-journal.com)


Stata Journal
Impact Factors
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1.433
1.522
1.851
2.000
2.222
2007
2008
2009
2010
2011
JCR Years
(ISI Web of Knowledge Journal Citation Reports)

Stata: an overview

- Some of the things I don't like about Stata:
  - Graphics look great and are quite flexible, but they are a bit slow
  - Only „traditional" maximum likelihood estimation, no general Bayesian algorithms
  - Support for nonparametric models could be better
  - Only one dataset in memory at the time (although you can have additional matrices and Mata objects)
  - String variables limited to 244 characters (although not in Mata), no UTF support
  - More tools for output processing and reporting would be valuable

Survey estimation

- One of the strengths of Stata is its support for complex surveys.
- This started early on in the history of Stata with the introduction of sampling weights and robust variance estimation for regression models.
- Through the `svy` prefix command, almost all estimation commands of Stata can be made „survey design aware“ (this also includes SEM, which appears to be a unique feature of Stata).
- The `svy` command supports
  - sampling weights
  - stratification
  - clustering
  - multistage sampling
  - finite population correction
  - post stratification

Survey estimation

- Supported methods of variance estimation are
  - Taylor linearization
  - jackknife
  - balanced repeated replication (BRR) and bootstrap
  - successive difference replication (SDR)

Survey estimation

- Supported methods of variance estimation are
  - Taylor linearization
  - jackknife
  - balanced repeated replication (BRR) and bootstrap
  - successive difference replication (SDR)
- Usage:
  1. Specify sample design using the `svyset` command.

     `svyset` *`psu_id`* `[pweight=`*`weight`*`], strata(`*`strata_id`*`)`

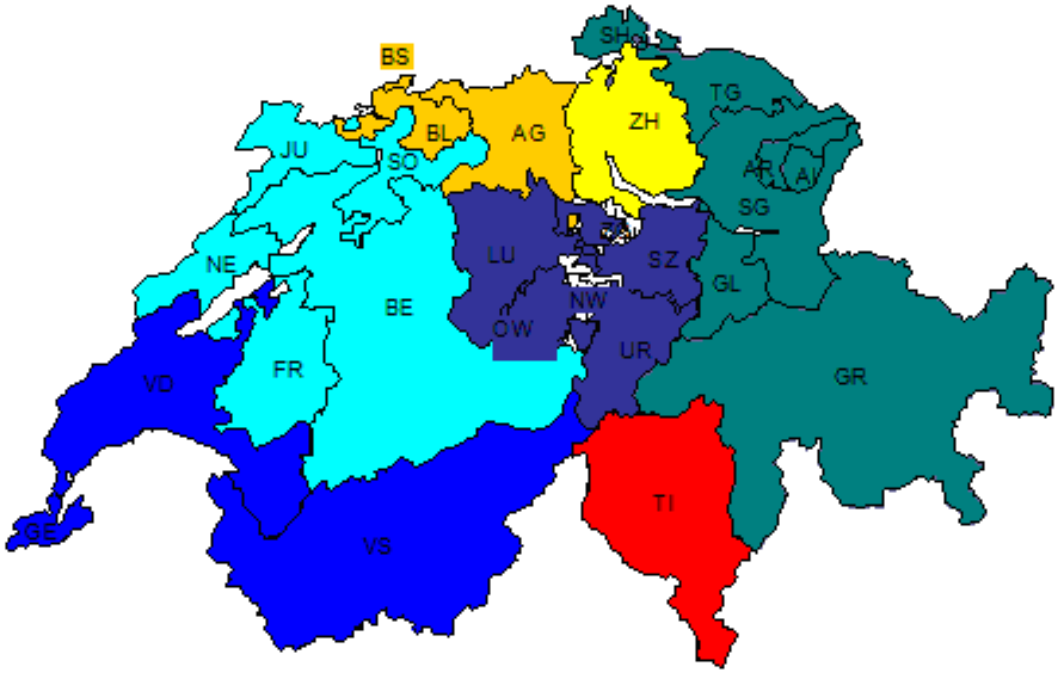
  2. When analyzing the data, apply the `svy` prefix.

     `svy:` *`command`*

Example: Swiss Household Panel (SHP)

- Stratification

Figure 1: The 7 major geographical regions of Switzerland (NUTS II regions)



(Plaza/Graf 2008)

- All household members are interviewed. That is, the sample is clustered by households (each household is a PSU)
- Sampling weights are used to compensate different selection probabilities, non-response, and panel attrition.

Predictive margins and marginal effects

- For a long time, regression tables have been the preferred way of communicating results from statistical models.
- However, interpretation of regression tables can be very challenging in the case of interaction effects, categorical variables, or nonlinear functional forms.
- Moreover, interpretational difficulties can be overwhelming in nonlinear models such as logistic regression. In these models the raw coefficients are often not of much interest; what we want to see for interpretation are effects on outcomes such as probabilities, not on „latent“ variables such as log odds.
- Fortunately, Stata has a number of handy commands such as `margins`, `contrasts`, and `marginsplot` for making sense of regression results.

Example: Factorial Survey on Just Incomes

- Mail survey among a random sample of the Swiss population ($N = 1945$). Written questionnaire in German, French and Italian.
- Respondents were asked to judge short text descriptions of (fictional) individuals (so called "vignettes"), in which c certain elements are varied at random.
- For our research objective, we used vignettes describing men and women employing the following $2 \times 2 \times 2 \times 3$ design :
  - male vs. female
  - single without children vs. married without children
  - average work effort vs. above-average work effort
  - income levels: 5000 CHF, 5500 CHF, 6000 CHF

The Vignette

In letzter Zeit wird viel über die Höhe von Löhnen in verschiedenen Berufen gesprochen. Wir interessieren uns für Ihre persönliche Einschätzung zu diesem Thema.

Stellen Sie sich die folgende Situation vor:

{Herr | Frau} Müller, 25-jährig, {allein stehend und ohne Kinder | verheiratet in kinderloser Ehe}, arbeitet als kaufmännische{r|} Angestellte{r|} im Rechnungswesen eines mittleren Dienstleistungsbetriebs und erbringt dort {überdurchschnittliche | durchschnittliche} Leistungen. {Sein | Ihr} monatliches Bruttoeinkommen beträgt {5'000 | 5'500 | 6'000} Franken.

Wie bewerten Sie das Einkommen dieser Person? Ist das Einkommen Ihrer Meinung nach gerecht oder ist es ungerechterweise zu hoch oder zu niedrig?

viel zu niedrig **gerecht** **viel zu hoch**

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

Thank you for your attention!